DOCUMENT RESUME

ED 310 138 TM 013 719

AUTHOR Sawyer, Richard
TITLE Validating the Use of Standardized Test Scores for Remedial Course Placement in College.
PUB DATE Mar 89
NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Entrance Examinations; College Freshmen; Decision Making; Grade Prediction; Grades (Scholastic); Higher Education; *Remedial Instruction; *Scores; *Standardized Tests; *Student Placement; *Test Validity
IDENTIFIERS ACT Assessment; *Placement Tests

ABSTRACT

This paper considers the problem of validating placement procedures or, more precisely, of determining their educational appropriateness. At issue is determining whether a test score serves the particular educational function it was designed to serve (for example, course placement), and whether it does so in an economical way. These determinations are made using an extension of the decision theory methods developed by L. J. Cronbach and G. C. Gleser (1965) and by N. S. Petersen and M. R. Novick (1976) in the context of recent ideas about validation. Methodological issues in quantifying system effectiveness are then examined and illustrated. The discussion covers types of placement selection rules, validating such rules, the plausibility of assumptions underlying M. Kane's (1987) validation paradigm associated with the use of entrance examination scores (American College Test) combined with high school grades, selection rules based on grade predictions, a decision model for validating placement rules, and determination of the effectiveness of remediation. An example is provided that is based on the American College Test scores, self-reported high school grades, and freshmen English course grades of students at a public university (N=5,609). Five tables and two figures are included. (TJH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED310138

# Validating the Use of Standardized Test Scores for Remedial Course Placement in College

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

**Richard Sawyer**

The American College Testing Program

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RICHARD SAWYER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting of the American Educational Research Association, March 30, 1989, San Francisco, California

The author thanks Michael Kane for his advice on the issues discussed in this paper, and Mark Houston, Jim Maxey, and Julie Noble for their comments on earlier drafts.

TM013719

BEST COPY AVAILABLE

## Validating the Use of Standardized Test Scores for Remedial Course Placement in College

A typical and important use of college admissions tests is college freshman course placement, i.e., the matching of students with appropriate instruction. For example, students with low predicted chances of success in a standard freshman English course might, on the basis of their test scores or other measures of their academic development, be advised or required to enroll in a remedial English course. On the other hand, students with high predicted chances of success in an accelerated English course might be encouraged to enroll in it. This paper considers the problem of validating placement procedures, that is, of determining their educational appropriateness.

Cronbach (1988) argued that validating a test use is really evaluation, and must not only address the particular educational functions for which the test was designed, but must also consider the broader educational, political, and social consequences of using the test from the perspectives of different value systems. By this standard, validation is an enormous and never-ending task. This paper is concerned only with the traditional issue in Cronbach's agenda, namely determining whether a test score serves the particular educational function it was designed to serve (e.g., course placement), and whether it does so in an economical way. For a discussion of procedures by which other aspects of placement systems can be evaluated, see Frisbie (1982).

An impressive theoretical methodology, based on statistical decision theory, has been developed during the past twenty-five years for determining the effectiveness of selection systems. A principal goal of this paper is to interpret the decision theory methods developed by Cronbach and Gleser (1965) and by Petersen and Novick (1976) in the context of recent ideas about

validation (Kane, 1987). Further, their decision models are extended to one that more easily addresses the concerns of an institution in measuring the effectiveness of its placement system. Methodological issues in quantifying system effectiveness are then examined, and are illustrated by an example.

## Remediation

Many different techniques exist for matching college students with appropriate instruction. Willingham (1974) described in detail and classified various placement techniques that are based on test scores and other measures of academic ability. This paper focuses on only one particular placement function, remediation.

At many postsecondary institutions, there are two levels of freshman courses: a standard level course in which most freshmen enroll, and a lower level course for students who are not academically prepared for the standard course. (At some institutions, there is a "developmental" or "review" course in addition to the remedial course; in this paper, we consider only the single remedial lower-level course.) In Willingham's classification of placement procedures, this instructional treatment is called "remediation" (Model 4).

A placement device, such as a test, is but one component of a placement system. To be educationally effective, the placement system must have at least all the following characteristics:

a. There must be some way to identify accurately those students who have a small chance of succeeding in the standard course.

b. Appropriate remediation must be provided to the high risk students.

c. Both the students who originally enrolled in the standard course, and the students who were provided remediation, must do satisfactory work in the standard course.

Note that merely accurately identifying high risk students is insufficient for the placement system as a whole to be effective. If these high risk students eventually drop out or fail in the standard course, no useful purpose will have been served by the placement system; on the contrary, both the institution's and the students' resources will have been wasted. The validation strategies described here address issues a. and c.

**Types of Placement Selection Rules**

An institution's placement rule is assumed to have the following general form: If a student is predicted to have a small chance of succeeding in the standard freshman course, then the student is selected for the remedial course. Thus, the strategy is to reduce the chances of failure by identifying students who are at high risk, then to offer them instruction at an appropriate lower level.

Placement is often based on a single test score related to a course. For example, placement in freshman English may be determined only from students' scores on an English placement test. A critical score level for selection is determined either by reviewing the contents of the test and college course, or by analyzing the statistical relationship between test score and course grade.

Single score cutoffs are also frequently determined on the basis of local normative information, the goal being that a predetermined number or percentage of students automatically be selected for each level. This has obvious administrative advantages in sectioning courses, and may be effective in correctly placing students, if the academic skills of students selected for

the different levels are appropriate. To ensure a match between students' skills and course requirements, however, the contents of the courses into which students are placed might need to be adjusted.

Alternatively, placement rules can be stated in terms of grade predictions or expectancies that are based on several predictor variables, such as test scores and high school grades. The advantage of using several prediction variables is, of course, that the predictions are potentially more accurate, and therefore more likely to result in correct placement decisions. For example, the American College Testing Program (ACT, 1987) predicts college freshman grades from students' ACT scores (in English, mathematics, social studies, and natural science) and from their self-reported high school grades in these subject areas.

It should be mentioned, as an aside, that some institutions select for an accelerated course or for advanced placement those students who have a very high chance of success in the standard level course. This procedure may work satisfactorily in practice, but it is not as directly linked to the placement goal as is the procedure described previously. The reason is that students who are predicted to do well in the standard level course could, nonetheless, be ill-prepared for the advanced course; this could occur if the skills required for success in the accelerated or advanced course were not measured by the placement test, and if they differed significantly from those required in the standard course. The practical advantage of this alternative placement method is that it requires developing only one prediction equation for all three course levels (remedial, standard, accelerated/advanced), rather than two prediction equations.

**Validating Placement Rules**

In validating tests for use in college admissions and placement, researchers have traditionally emphasized documenting time-ordered statistical relationships between test scores and relevant criteria. Typically, this documentation has consisted of correlation coefficients and associated tests of statistical significance. To the extent that the usefulness of a test depends on the existence of statistical relationships, such evidence is relevant to validation. There are advantages however, to moving beyond simple documentation of validity statistics to a more theoretically oriented validation strategy. In the more theoretical approach to validity, the use of a test for a particular purpose is seen as the logical conclusion of a set of assumptions which need to be justified; the statistical analyses usually thought of as constituting the validation become the means by which some of the assumptions can be justified, either directly or indirectly. Such a procedure can be thought of as applying the scientific method to test validation. Angoff (1988, p. 30), in summarizing the trend during the past three decades to think of validation this way, stated, "...as it became clear that it was not the test, but the inferences drawn from the test scores that were on trial, it also followed that the theory that dictated the inferences was also on trial."

Purely empirical validation through extended replication and variation can, within a limited sphere, make certain claims credible and certain predictions safe to use---much of early human knowledge, for example, probably developed atheoretically from millenia of trial-and-error validation. A theoretical approach to validity, however, offers the possibility of understanding why empirical relationships exist, of making educated guesses about what limitations there may be in extrapolating them to different

situations, and even of figuring out how new tools might be devised that are more effective in accomplishing our goals.

Kane (1987) proposed a paradigm for validating uses of test scores that is theoretically oriented. In Kane's approach to validation, one first states as clearly as possible the particular use being made of test scores, and the logical chain of assumptions by which the use can be justified. Next, one examines the plausibility of each of the assumptions. One then investigates more thoroughly those assumptions which are least plausible, based on available evidence; usually, this will involve collecting and analyzing relevant data. The final step is to review the overall plausibility of the logical chain of inferences, and to determine how the plausibility can be enhanced, either by modifying the test or the use made of the test. Although Kane proposed this method in the context of professional licensure and certification, it is easily transferred to other contexts, including course placement.

Clearly, Kane's approach to validation would require a different validity argument for different tests or different uses of the same test; presumably, different arguments could also be made for the same test and the same use. Following is a simple argument for using achievement-oriented tests (like the ACT Assessment) and self-reported high school grades to identify high risk students. It is based on two assumptions:

1. The academic skills a student acquires from taking a college course are directly related, among other things, to the academic skills the student has previously acquired. Moreover, there are minimum skills the student must bring to the course before he or she can be expected to derive much benefit. The skills required are particular to each college course: they may overlap to some extent (e.g., reading skills are necessary for

most courses), but also have unique elements (e.g., knowledge of analytic geometry is a prerequisite for calculus, but not for any English course).

2. The test scores and self-reported high school grades provide either direct measurements or indirect indicators of the required skills.

Note that no claim is made that prior educational achievement is the only determinant of student performance in college, or the only practical basis for placing students. Other student characteristics could also be important (or conceivably, more important) and could be included in the validity argument by making additional assumptions. The simple argument described here is a foundation on which to construct a justification for using achievement-oriented tests for placement. A different argument would, of course, be required for using aptitude-oriented tests for placement.

**Plausibility of Assumption 1**

It is difficult to conceive of any college level course in English, mathematics, social studies, or natural science for which the first assumption is not true; indeed, the structure of all educational systems seems to take this as a given. If, though, a college-level course of the type we are considering did not require previously acquired academic skills, then there would be no need for placement. More typically, the need for a placement system results from a practically significant number of students not possessing these skills.

**Plausibility of Assumption 2 (ACT test scores)**

Before using achievement-oriented tests for placement, college staff should review their contents to determine their relationship with that of the standard college course. Following is a general discussion of the contents of the ACT tests (ACT, 1987) and their relationship to typical college courses.

The ACT Assessment cognitive tests were designed to measure directly the academic skills needed to do college-level work in English, mathematics, social studies, and natural science. The tests are oriented toward the content of secondary and postsecondary programs in these fields, rather than toward a factorial definition of various dimensions of intelligence, and are intended to have a direct and obvious relationship to students' academic development. The tasks included in the tests are representative of academic skills typically taught in high schools and needed for postsecondary work; they are intended to be comprehensive in scope and educationally significant, rather than narrow or artificial. They rely partly on students' reasoning abilities and partly on their knowledge of the subject matter fields, and emphasize the use of reasoning skills in conjunction with knowledge of subject matter.

It is unlikely that the ACT Assessment (or any other battery of multiple choice tests with similar content, length, and breadth) measures all the academic skills required for a particular college course. It is likely, though, that for many freshman courses the ACT Assessment directly measures many of the important required skills. Because better prepared students will probably learn more in a college course than less well prepared students (Assumption 1), one can reasonably expect that ACT test scores would be statistically related to students' performance in such courses.

In addition, some college courses may require skills that are not directly measured by the ACT Assessment, but that are closely related to skills the ACT Assessment does measure. A good example of this is writing skills, which are obviously necessary for many college courses. The ACT Assessment does not provide a direct measurement of students' writing skills, as for example, from a writing sample. The ACT English Usage test does,

however, measure students' skills in editing, which are closely related, both conceptually and statistically, to their writing skills.

**Plausibility of Assumption 2 (High School Grades)**

High school grades are another traditional measure of students' readiness for college level study. To use them for placement, college staff should, at least in principle, review the contents of individual courses at particular high schools (as they would the contents of the placement test) to determine their relevance to the college course. In practice, this is not usually feasible; but, if the assumed contents of the high school courses required for admission to the institution have a plausible relationship to the college course, then one may reasonably assume that students' grades in these courses will be related to their readiness for the course.

High school grades probably also measure students' socialization, motivation, work habits, and study skills, as well as their academic skills and knowledge. In the simple validity argument based solely on prior achievement, these other factors are irrelevant. On the other hand, the amount students learn in a college course may well be related to these factors, and high school grades may provide a broader perspective on the likely benefit to students in taking the course. In this case, the validity argument could be expanded to take into account factors other than achievement.

In the ACT Assessment, high school grades are self-reported. Sawyer, Laing, and Houston (1988) compared the grades reported by students in 30 standard high school courses with corresponding grades from the students' transcripts. They found that in typical courses, 71% of students reported their grades exactly; and that 97% reported their grades accurately to within 1 grade unit. Moreover, specific course grade predictions based on self-

reported high school grades are almost as accurate as predictions based on ACT test scores; and predictions based on both high school grades and test scores combined are more accurate than predictions based on either alone (Noble and Sawyer, 1987). One can conclude that some individuals' self-reported grades may be of doubtful accuracy, but that on the whole, self-reported grades can be useful in placement.

**Selection Rules Based on Grade Predictions---An Additional Assumption**

Determining the content "fit" among test scores, high school grades, and a particular course at a given postsecondary institution must, of course, be done by individuals who know the course content at the institution. If the fit of the tests and high school courses to the college course is good, it is reasonable to expect that students with higher test scores and high school grades will outperform students with lower test scores and high school grades. It is therefore appropriate to consider using these two kinds of measures for course placement.

Practical implementation of a course placement procedure requires that some decision rule be formulated in terms of critical values of the test scores and high school grades. In principle, one could determine the critical values on the basis of expert judgment about test content and high school curriculum, and about their overlap with the content of specific college courses. Jaeger (1989) provides an overview of standard setting procedures based on judgments. Making such judgments would likely be very difficult, especially if the placement rule were based on more than one variable.

The easiest and most common way to implement a placement decision rule based on multiple measures of ability is through a prediction equation for course grades. Placement decisions are made on the basis of a critical value, either for the predicted grade or for its transformation to a grade

expectancy. Moreover, measures of the strength of the statistical relationships between placement variables and course grades are routinely provided by the software that calculates prediction equations; these summary measures provide additional support of the plausibility of Assumption 2. The appropriateness of grade predictions, however, depends on an additional assumption:

3. The course grades being predicted are valid measures of the academic skills learned in the course, rather then measures of random or irrelevant factors.

If this assumption is not met, then it would be inappropriate to base placement decisions on grade predictions, and data on the statistical relationship between placement variables and course grades would be irrelevant to the validity argument. Of course, if this assumption is not met, an institution has a much more serious problem than validating placement rules!

Lesser reliability in the course grade will (other things being equal) always result in lesser prediction accuracy, as indicated by smaller multiple correlations and larger standard errors of estimate. Assuming that one has determined that the predictor variables are related to course content, then a reliable but invalid course grade would also generally result in smaller multiple correlations and larger standard errors.

One could also hypothesize a situation in which the predictors and a highly reliable course grade were all unrelated to mastery of course content, but were related to the same irrelevant factor. In this situation, course grade prediction equations could have respectable multiple correlations and standard errors, but still be inappropriate for use in placement.

Following is an example of potential "shared invalidity". Both high school grades and college course grades are ratings by individual instructors,

and both may be influenced by factors other than their students' academic accomplishments. For example, some instructors may be more likely to give higher grades to students who attend every class, who turn in assignments punctually, or who are courteous, well-groomed, and appear to be interested in what the instructors have to say, than they are to students who are not so well socialized. Such grading practices may or may not be appropriate, depending on the values of the instructor and the policies of the high school and college, but to the extent that they do occur, the statistical relationship between high school grades and college course grades will be made less relevant to validating a placement rule based on educational achievement.

The reliability and validity of course grades must, of course, be determined at each individual institution. How to make this determination is a difficult psychometric problem, and is beyond the scope of this study. The few published results available are briefly summarized here. Etaugh, Etaugh, and Hurd (1972) obtained an estimated single course grade reliability of .44 using the single course grades for each student as repeated measurements. Schoenfeldt and Brush (1975) adapted the procedure of Etaugh, et al. to estimate reliabilities in 12 different course areas; their estimates ranged from .39 to .76. Although both estimation procedures may be open to question, the results suggest that specific course grades are not as reliable as ACT test scores, for which parallel form reliabilities approaching .9 are typically obtained (ACT, 1987). Assuming a reliability of .9 for ACT tests and .4 for specific course grades, one cannot expect correlations between single ACT test scores and course grades to exceed .6.

Duke (1983) found that the distribution of grades awarded at a particular university varied markedly by department, and that students who earned high grades in departments with easy grading earned average or low grades in other

departments with stringent grading. He concluded that GPA is made up of components that are not equivalent among students. His findings also suggest that grades in some departments measure, to a significant extent, characteristics other than students' academic achievement. Although Duke investigated the grading practices of only one institution, he believed his results were typical of those in higher education generally. I shall assume, for the purpose of this discussion, that the course grade being predicted measures academic achievement, and is therefore relevant to the goals of the placement function. Duke's results indicate, though, that this assumption needs to be justified in particular applications.

**A Decision Model for Validating Placement Rules**

Given that the contents of the placement test and high school courses are reasonably (but less than perfectly) congruent with the skills required in the college course, and that the course grade is reasonably (but less than perfectly) reliable and valid, one can expect there to be a (less than perfect) statistical relationship between test scores, high school grades, and college course grades. This expectation can be tested against data collected by the institution, and if not borne out, would lead one to reconsider the plausibility of the assumptions in the validity argument.

But, assuming that the expectation of a statistical relationship is borne out, how does one quantify the usefulness of placement decisions based on the grade predictions made possible by this relationship? The answer, in general, is that the validity argument must be augmented with additional assumptions about the benefits of student achievement and the costs of providing instruction. These assumptions would need to address all the important outcomes of the placement system, such as the performance of students who are

placed in the remedial course, as well as that of students who are placed in the standard course. These assumptions can then be related to the statistical relationships estimated from the data to produce a summary of the usefulness of the placement system as a whole. Note that at this point we are making inferences about an entire placement system, of which the placement test is but one component.

Statistical decision theory has been proposed by several writers, including Cronbach and Gleser (1965) and Petersen and Novick (1976), as a useful means for analyzing educational selection problems. Validating placement systems, in their full generality, through decision theory is a complex undertaking. To structure the discussion, let us first consider the requirement that a placement system accurately identify high risk students.

Suppose we could collect test scores, high school grades, and course grades for a representative sample of students who have all taken the standard course without any prior remediation. Some of these students would be successful, as measured by their course grades; others, presumably, would be unsuccessful. The students' predicted and actual grades could then be compared and a numerical value could be assigned to each outcome.

**Table 1**
**Outcomes Among a Group of Students**
**Who Take the Standard Course**

| Performance in standard course | Predicted grade: Below critical value | Predicted grade: Above critical value |
|---|---|---|
| Successful | D | A |
| Unsuccessful | C | B |

An example of a simple decision model is given in Table 1, in which there are four possible outcomes. Outcome A is called a "true positive"; outcome B, a "false positive"; outcome C, a "true negative"; and outcome D, a "false

negative". Let f(A), f(B), etc., denote the frequencies for outcomes A, B, etc. in the group of students. Then f(A) + f(C) is the number of students for whom correct placement decisions would have been made had the placement procedure been used; and f(B) + f(D) is the number of incorrect decisions. The overall usefulness of the predictions would then be evaluated in terms of the benefits of correct placement decisions and the losses resulting from incorrect placement decisions. A function that assigns a value to outcomes such as these is called a utility function. For this model, in which outcomes for groups of students are considered, a simple type of utility function would be the frequency or proportion of correct placement decisions; according to such a utility, every correct decision results in the same positive benefit and every incorrect decision results in zero benefit. A more complex utility function would assign different values to each outcome: $k_A f(A) + k_B f(B) + k_C f(C) + k_D f(D)$, where $k_A, k_B, k_C, k_D$ are constants. Such a function would quantify the different benefits of true positives and true negatives, as well as the different costs of false positives and false negatives.

The utility function of this model superficially resembles the expected utility function (see below) in the decision model developed by Petersen and Novick (1976). Their "threshold" utility function, however, pertained to the possible outcomes for an individual student, rather than for a group of students. By considering outcomes for groups of students, this model can directly address an institution's utilities for the results of its placement system. For simple utility functions like the one described above, in which an institution's utility is the sum of utilities for individual students, the two approaches amount to the same thing. The group model, however, by considering placement outcomes from an institutional, rather than individual perspective, permits one to develop utility functions that address the more

complex outcomes that an institution, rather than an individual, must consider.

For example, an institution's utility function should consider the costs of providing remedial instruction to high-risk students. With each additional student placed in the remedial course, one can expect a small increment in cost until a new instructor must be hired and an additional classroom provided. At this point, the total cost jumps by a large amount, and the per-student cost must be recomputed. Therefore, an institution's utility function can not be represented as a simple sum of utilities for individual students; it must take into account the total number of students assigned to remedial instruction. One might approximate such a step function by a linear function, which could be used in the individual level model, but the structure of the decision problem is more easily conceived in the group model.

An institution's utility function should also take into account the cost of testing. For example, a placement rule could be based on using a general purpose, relatively low cost battery like the ACT Assessment for most students, and a custom designed, higher cost, local placement test for students with ACT scores near the critical score level. Such a placement rule could be more advantageous to an institution than one based on either test alone, when both costs and accuracy of placement are considered. Cronbach and Gleser (1965) discussed two-stage testing in the context of a utility function that is assumed to be linearly related to the scores on the two tests and that can be related to the cost of testing. They showed how, given these assumptions, the critical scores on the two tests and the resulting gain in efficiency can be determined.

A difficult practical problem in implementing decision models is relating an institution's costs to its students' performance. Essentially, an

institution must determine how much, in dollars, any given level of student performance is worth. Although institutions implicitly make such judgments whenever they decide on how to allocate their resources, explicitly declaring their values by way of a utility function is difficult, both technically and politically. (Cronbach and Gleser called it the "Achilles' heel of decision theory".) Reilly and Smither (1985) studied various methods that have been proposed in employment testing for linking criterion variables to a dollar scale. They found the methods effective in artificial situations where much information was available, but they were less sure about how effective the methods would be in more general situations. Clearly, much work needs to be done before utility functions can be routinely developed for educational testing applications like placement.

An institution's utility function reflects its concerns, which do not necessarily coincide with those of individual students. On the other hand, the institution must take care in formulating its utility function that the rights of individual students are respected; in particular, the resulting placement policies should be consistent with the Standards for Educational and Psychological Testing (APA, 1985). Whitney (1989) described ways in which institutions' and students' utilities typically differ.

Expected Utilities

In evaluating a placement rule with respect to a utility function applied to the outcomes of future students, the frequencies f(A), f(B), etc., are not known. These frequencies must instead be estimated under the assumption that future students will be like, at least in some ways, the students for whom there are data. A utility function like the one considered here, when expressed in terms of estimated cell frequencies or proportions, is called an

expected utility; it is from the expected utility function that decisions on the effectiveness of a placement rule can be made.

More precisely, an expected utility in a Bayesian decision model is the mean of a utility function with respect to the subjective probability distribution of the unknown parameters in the function. In the model described by Table 1, the unknown parameters are the frequencies f(A), f(B), etc. for future students. These parameters could be modelled directly with a multinomial distribution, or, alternatively, they could be inferred from a model of the joint density of course grades and test scores. In the example given later in this paper, statistical inferences are grounded in classical (frequency) probability, but Bayesian (subjective) probability models are being developed (Houston, 1988). The merits of alternative statistical techniques for computing expected utilities, while important, are not considered in this paper. Rather, I have chosen to minimize issues of statistical inference and to emphasize how expected utilities can be used to evaluate placement rules.

Note that expected utilities can be used not only to evaluate an existing critical score for placing students in the remedial course, but also to determine the optimal critical score. In the model described by Table 1, for example, each potential critical score is associated with a potentially different expected utility; the critical score level can be selected that maximizes the expected utility.

Clearly, the components of a utility function, such as the benefits of correct placement decisions and the costs of incorrect ones, vary among institutions. Furthermore, the statistical characteristics of grades and test scores will be unique to each institution. Therefore, computing an expected

utility requires the involvement of local institutional staff, as does constructing the other components of a validity argument.

**Determining the Effectiveness of Remediation**

With an appropriate placement rule, students who are likely to be unsuccessful in the standard course can be identified and placed in a remedial course. Although one reduces the likelihood of immediate failure by such intervention, the question remains whether students placed in the remedial course will later succeed in the standard course. To extend the validity argument previously discussed, one then needs to examine the plausibility of the following additional assumption:

4. The remedial course provides, in a cost-effective way, the academic skills that students previously identified as high risk need to succeed in the standard course.

To establish the plausibility of this assumption, it is clearly appropriate to examine the syllabus for the remedial course: There should be a fit between the remedial course contents and the academic skills needed to succeed in the standard course, as previously identified in justifying Assumption 2.

Given the other assumptions in the validity argument, it is also appropriate to examine separately for remedial course students the relationship between their placement test scores and the grades they finally earn in the standard course. If this relationship is the same as that for students who enrolled directly in the standard course, then the remedial course is of no benefit---students with similar predicted grades from the placement test have the same outcome (i.e., they tend to earn unsatisfactory grades in the standard course), whether or not they take the remedial course. For the placement system as a whole to be successful, the students

placed in the remedial course should, on average, have higher grades in the standard course than they would have if they had not been placed in the remedial course. Note that this concept relates to the placement system as a whole, and not just to one component (such as the placement test).

If a decision model is to be an effective part of the validity argument concerning remediation, then it must consider the costs, as well as the benefits, of differential treatment. The utility function in such a model would place a value on the ultimate performance of every student in the standard course, and it would take into account the extra cost incurred when students first enroll in the remedial course. An institution would need to verify that low scoring students who enroll in the remedial course have a higher expected utility than do low scoring students who enroll directly in the standard course. The institution would also need to verify that high scoring students who enroll in the remedial course have a lower expected utility than do high scoring students who enroll directly in the standard course. This relationship between expected utility, test score, and treatment is an example of a statistical interaction, and is illustrated in Figure 1. Note that the vertical axis in this figure is a utility function that takes into account the extra cost associated with providing additional instruction to the remedial students.

If the vertical axis in Figure 1 represented the standard course grade instead of a utility function, then the two lines might not intersect or might intersect at a different point. Cronbach and Snow (1977, pp. 32-33) provide a discussion and illustration of this phenomenon. Moreover, if the regression slope of course grade on test score is the same for both treatment groups (and if utility is linearly related to course grades), then there can be no interaction in the regression of utility on test score. The reason is that

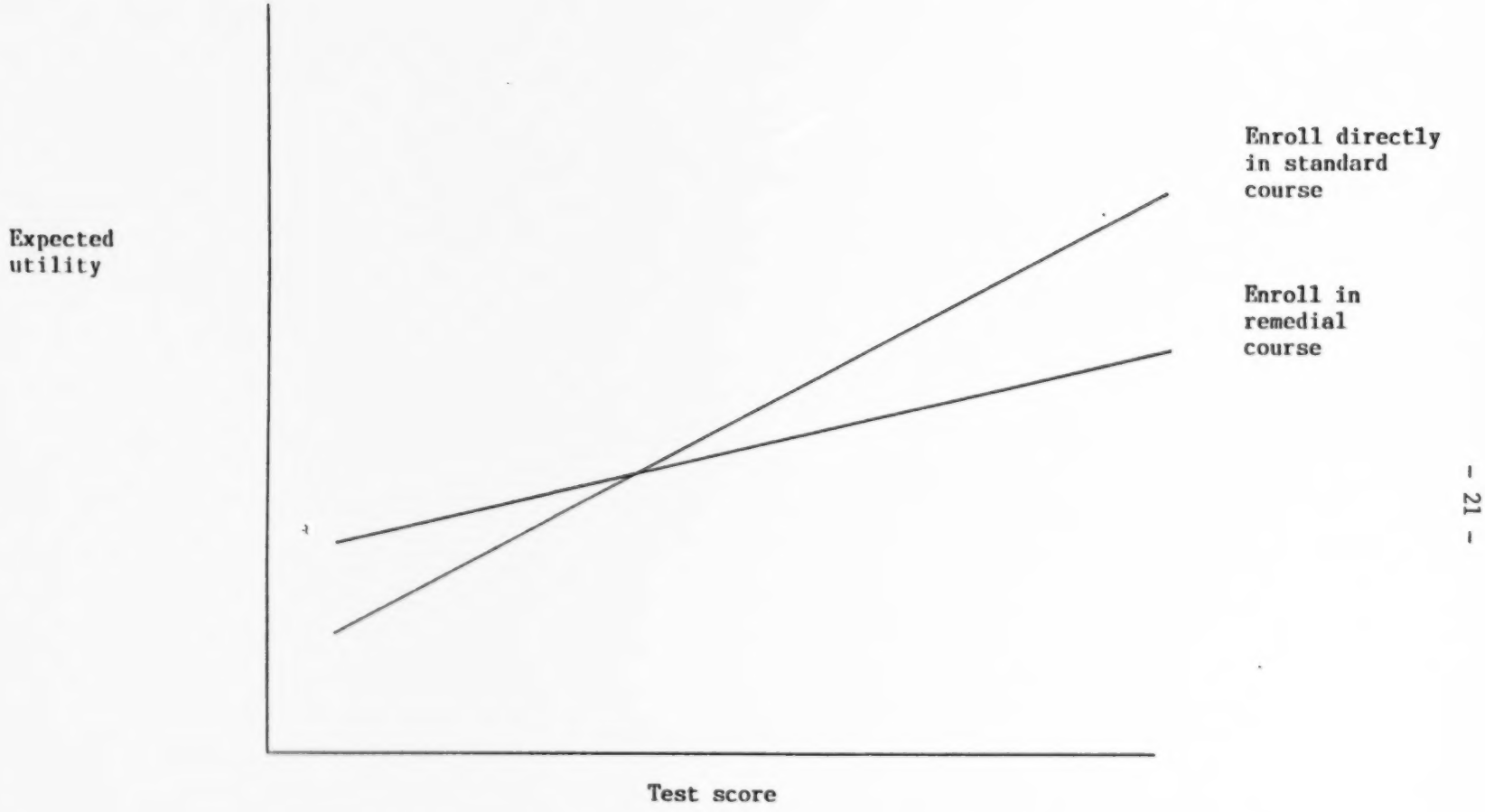


Figure 1. Relationship Between Expected Utility Function and Test Score, by Treatment Group

the cost of providing remedial instruction is not a function of test score; therefore, the regression line for utility can differ from the regression line for course grade in intercept only. If the regression slopes for the two groups' course grades are equal (or nearly equal), then the regression lines for their utility functions will not intersect, and one treatment will always be judged superior to the other. Therefore, the slopes of the regression functions for course grades can be used to make inferences about the presence of interactions in the utility function regressions.

In theory, treatment effects can be determined by randomly allocating students either to the remedial course (treatment group) or the standard course (no-treatment group), then studying the relationship between standard course grade (or utility) and test score for the two groups. In practice, students are not randomly allocated to the remedial and standard courses, and this considerably complicates proper interpretation of treatment effects. At an institution where there is an existing placement system, low scoring students do not take the standard course before they have completed the remedial course; and high scoring students do not take the remedial course. Therefore, such an institution can ordinarily estimate the relationship between course grade (or utility) and test score for the no-treatment group only in the upper range of test scores. In an institution without a validated placement system, it might be possible to assign low scoring students randomly to the treatment and no-treatment groups until the system could be validated. Otherwise, estimates of treatment effects must usually be based on the extrapolation of the no-treatment group regression to the lower range of scores (Cronbach and Snow, 1977, pp. 48-49).

When the placement test is but one component of the placement decision, as in a voluntary system, there will be greater variation in the test scores

of each group, and the need for extrapolation would appear to be reduced. Unfortunately in this situation, the differences in the regression lines are confounded with whatever variables were used to make the placement decision. It is impossible to conjecture what effects the confounding has, in general, though it might be possible to do so in a particular placement system. For example, if the other variables can be quantified, they can be incorporated into a model with test score, treatment effect, and score by treatment interaction.

**Example**

This example is based on the ACT scores, self-reported high school grades, and freshman English course grades of students who enrolled in a medium size public university between summer, 1985 and fall, 1986. This institution encourages students with ACT English Usage scores between 1 and 15 to enroll in a remedial course, those with scores between 16 and 25 to enroll in a standard course, and those with scores of 26 and higher to enroll in a more advanced course. Placement is not determined solely by this rule, however; some students enroll in courses at a higher or lower level than recommended, and some students do not even have ACT scores. To simplify the analyses and the discussion, I retained only records with complete data and with enrollment patterns consistent with the placement rule just described. (Of 6,356 records in the total data set, 5,609 (or 88%) met these criteria.) For students who enrolled in the remedial course, both the grade in the remedial course and the eventual grade in the standard course were recorded. Records of students who enrolled in the advanced course were not analyzed.

According to the catalog for this institution, the standard English course teaches students to explore and communicate ideas through writing, and

emphasizes various prose patterns and techniques. The ACT English Usage test measures students' skills in the mechanics and style of written English. The test consists of passages with certain segments underlined; students are then asked, in multiple choice items, to indicate whether or how the underlined segments could be improved. The contents of the test would therefore appear to be relevant to the requirements of the course, and Assumption 2 of the validity argument would be plausible.

With regard to Assumption 3, one would need to know what aspects of students' performance were graded, and whether different instructors used the same grading standards. Unfortunately, no information was available on the grading methods used, and therefore, the plausibility of Assumption 3 can not be readily determined. I shall assume, in order to continue the discussion, that the grades predominantly measure aspects of students' performance related to the academic skills acquired in the course, and that different instructors had consistent grading standards.

Summary statistics for the ACT English score and the English course grades of the remedial and standard groups are given in Table 2. Note that 736 of the original 951 students in the remedial course persisted through the standard course, and their average grade in the standard course was .45 grade units lower than the average grade of the 4,463 students who initially enrolled in the standard course.

Table 3 contains statistics associated with the prediction of the standard course grade for the 4,463 students who initially enrolled in the standard course. Statistics are presented for predictions based on ACT English score alone (Model I) and for predictions based on the four ACT scores and four high school grades (Model II). Moreover, statistics were calculated both from the total group of cases with valid values and from a truncated data

Table 2

Summary Statistics for ACT English Scores and English Course Grades, by Placement Group

| Placement group | N | ACT English score | | Remedial course grade | | Standard course grade | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | Mean | SD |
| Remedial | 951 | 12.3 | 5.9 | 2.74 | .87 | 2.43[a] | .72[a] |
| Standard | 4463 | 20.4 | 6.0 | ---- | --- | 2.88 | .76 |

[a] Based on N = 736 records.

Table 3

Regression Statistics Associated with Predicting
End of Term Grade in Standard English Course

| Model | Statistic | Data set: Total group | Data set: Outliers deleted |
|---|---|---|---|
| I | N | 4463 | 4318 |
| (1 predictor) | Regression Coefficients | | |
| | Intercept | 1.15 | 1.43* |
| | ACT English | .085* | .075* |
| | R | .27 | .29 |
| | SEE | .73 | .61 |
| II | N | 3951 | 3793 |
| (8 predictors) | Regression coefficients | | |
| | Intercept | .64 * | .85 * |
| | ACT English | .057* | .056* |
| | ACT Mathematics | .002 | .003 |
| | ACT Social Studies | .006 | .008* |
| | ACT Natural Science | - .008 | - .007* |
| | HS English | .155* | .136* |
| | HS Mathematics | .059* | .035* |
| | HS Social Studies | .099* | .075* |
| | HS Natural Science | .040 | .040* |
| | R | .38 | .40 |
| | SEE | .70 | .60 |

* Denotes regression coefficients significant at the .01 level.

set with outlier cases removed. Outliers were defined as those for which either the studentized residual or the leverage statistic (Belsley, Kuh, and Welsch, 1980) exceeded the 99th percentiles of their distributions, assuming a normal distribution of errors. Of course, any inferences based on the truncated data set pertain only to the non-outlier portion of the population.

Note in Table 3 that Model II had larger correlations and smaller standard errors of estimate than did Model I. An anomalous result in Model II is the presence of a negative regression weight corresponding to ACT Natural Sciences; fortunately, the magnitude of this coefficient is not large enough to have much practical effect on the predicted English grade. Finally, note that removing the outlier cases tended to increase the correlations only slightly, but considerably reduced the standard errors of estimate.

The correlations associated with the truncated data sets (.29 and .40) were adjusted for prior selection, using standard procedures (Lord & Novick, 1968). In the calculation for Model I, the adjusted correlation was .41. For Model II, ACT English (X) was considered the explicit selection variable and the predicted English course grade (YHAT) was considered to be a proposed selection variable; the resulting adjusted correlation was .55.

Table 4 contains estimated cell probabilities associated with the "pass" criteria of C or better and B or better. Two of the estimates are based on the assumption that (X, Y) and (YHAT, Y) have approximate bivariate normal distributions with the adjusted correlations just derived. Observe that the estimated hit rate for C or better using placement based on Model I is .78 + .02 = .80, and the corresponding hit rate from Model II is .81. In this particular example, placement based on the single variable ACT English is almost as effective as placement based on all four ACT scores and all four high school grades. Placement based on eight variables was, however, somewhat

Table 4

Estimated Cell Probabilities Associated with
Grades in Standard English Course

| "Pass" criterion | Outcome | Prediction model: Bivariate normal, 1 predictor | Prediction model: Bivariate normal, 8 predictors | Prediction model: Logistic, 1 predictor |
|---|---|---|---|---|
| C or better | True positive | .78 | .78 | .80 |
| | True negative | .02 | .03 | .02 |
| | False positive | .02 | .02 | .03 |
| | False negative | .18 | .17 | .15 |
| B or better | True positive | .59 | .61 | .61 |
| | True negative | .11 | .13 | .12 |
| | False positive | .21 | .19 | .21 |
| | False negative | .09 | .07 | .06 |

more effective than placement based on ACT English alone, as judged by the hit rates for B or better (.74 versus .70).

Note also that using ACT English as a placement test reduced the failure rate, as determined by the standard of C or better, from .04 in the entire population to .02 in the selected population, though at the cost of a false negative rate of .18. By the standard of B or better, the failure rate was reduced from .32 to .21, and the false negative rate was .09.

Table 4 contains another set of estimated cell probabilities associated with placement using the ACT English test. These estimates are based on logistic regression, which is in several ways more direct and straightforward than the linear regression/normal theory methodology discussed so far. In logistic regression, a dichotomous outcome (such as C or better) is modeled directly from the predictor variable:

(1) $P[W=1 \mid X=x] = 1/[1 + \exp(-a-b*x)]$,

where W=1 if a student's grade exceeds a certain threshold, and W=0 otherwise. Using iterative estimation techniques (such as Gauss-Newton), it is possible to find the constants a and b for which the fitted conditional probabilities are closest to the observed outcomes, in the sense of weighted least squares. In general, it is more difficult to compute parameter estimates for a nonlinear model like this one than for a linear model. Once parameter estimates have been computed, though, estimated cell probabilities can easily be obtained by averaging the fitted conditional probabilities over the relevant values of x. For example, the rate of true positives can be estimated as:

$$(2)\quad \hat{P}\,[W=1 \mid X > 16] = \frac{\sum_{x=16}^{33} \hat{P}[W=1 \mid X=x]\, n(x)}{\sum_{x=16}^{33} n(x)}$$

where $\hat{P}\,[W=1 \mid X=x]$ is the conditional probability estimated from the logistic model (1), and n(x) is the number of observations of x. One could also assume a statistical model for the marginal distribution of X, and adapt the empirical frequencies n(x) accordingly (W. M. Houston, personal communication, 1989).

Estimates of the parameters a and b in equation (1) were computed using the SAS LOGIST procedure (SAS, 1986). The estimates were -3.36 and .24, respectively, for B or better; for C or better, they were -.36 and .19, respectively. Both estimated b coefficients were statistically significantly different from 0 ($p < 10^{-4}$).

The cell probabilities estimated from the fitted logistic curves are shown in the right-most column of Table 4. Note that they are very similar to the cell probabilities based on adjusted correlations and the bivariate normal distribution. This result illustrates that while the correlation coefficient can be made relevant to placement validity issues, given certain assumptions, it is not essential.

Estimated B-or-better cell probabilities were also computed for several hypothetical critical scores near the actual critical score of 16 (described previously). The hit rates associated with these estimates provide one means of judging the suitability of different critical scores. It turned out that the largest estimated hit rate (.73) was associated with the actual critical score of 16, although the critical scores of 15 and 17 had hit rates very

Table 5
Statistics Related to the Regression of English Course Grade
on ACT English Score, by Treatment Group

| | Group | N | Regression statistics for predicting grade in standard course[a] | | | | Average predicted grade | | Estimated proportion C or better | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Intercept | Slope | R | SEE | Group A equation | Group B equation | Group A equation | Group B equation |
| A. | Treatment (Remedial) | 863 | 1.86 | .049 | .75 | .63 | 2.53 | 2.47 | 1.00 | .97 |
| B. | No treatment (Standard) | 4553 | 1.46 | .073 | .32 | .61 | ---- | ---- | ---- | ---- |

[a] All estimates are based on data sets with outliers deleted.

nearly as large. Thus, when judged by the standard of the hit rate for B or better, 16 is the optimal critical score.

Finally, let us consider the issue of treatment effects in the remedial course. Table 5 contains statistics for the regression of standard course grade on ACT English score. These statistics are based on the records both of students who first enrolled in the remedial course (Group A) and of students who enrolled in the standard course to begin with (Group B). Table 5 is based on the full data set, rather than on the edited data set used to compute Tables 3 and 4, so as to reduce the amount of extrapolation. Outliers were deleted from the full data set according to the procedures followed in calculating correlation coefficients. The difference in the regression slopes for Groups A and B is statistically significant ($p < .001$), and allows the possibility of an interaction effect in the regression of the utility function on test score.

Results are also displayed graphically in Figure 2. The thick solid lines in the graphs pertain to predicted grades calculated over the central 80% of the cases; the thinner solid lines pertain to the predicted grades of the outermost 20% of the cases; and the dashed lines indicate extrapolations of the predictions to the remainder of the ACT English Usage test score scale. Note that the lines for the remedial and standard course students intersect near the ACT English score of 16, indicating an interaction. If the vertical axis in Figure 2 were an expected utility that incorporated the extra costs associated with providing remedial instruction, then the intersection point would have been at a lower ACT English score.

Two other statistics, pertaining to the students in the remedial course, are shown in Table 5. One statistic is the average predicted grade, and the other is the proportion of students whose predicted grades are 2.0 or

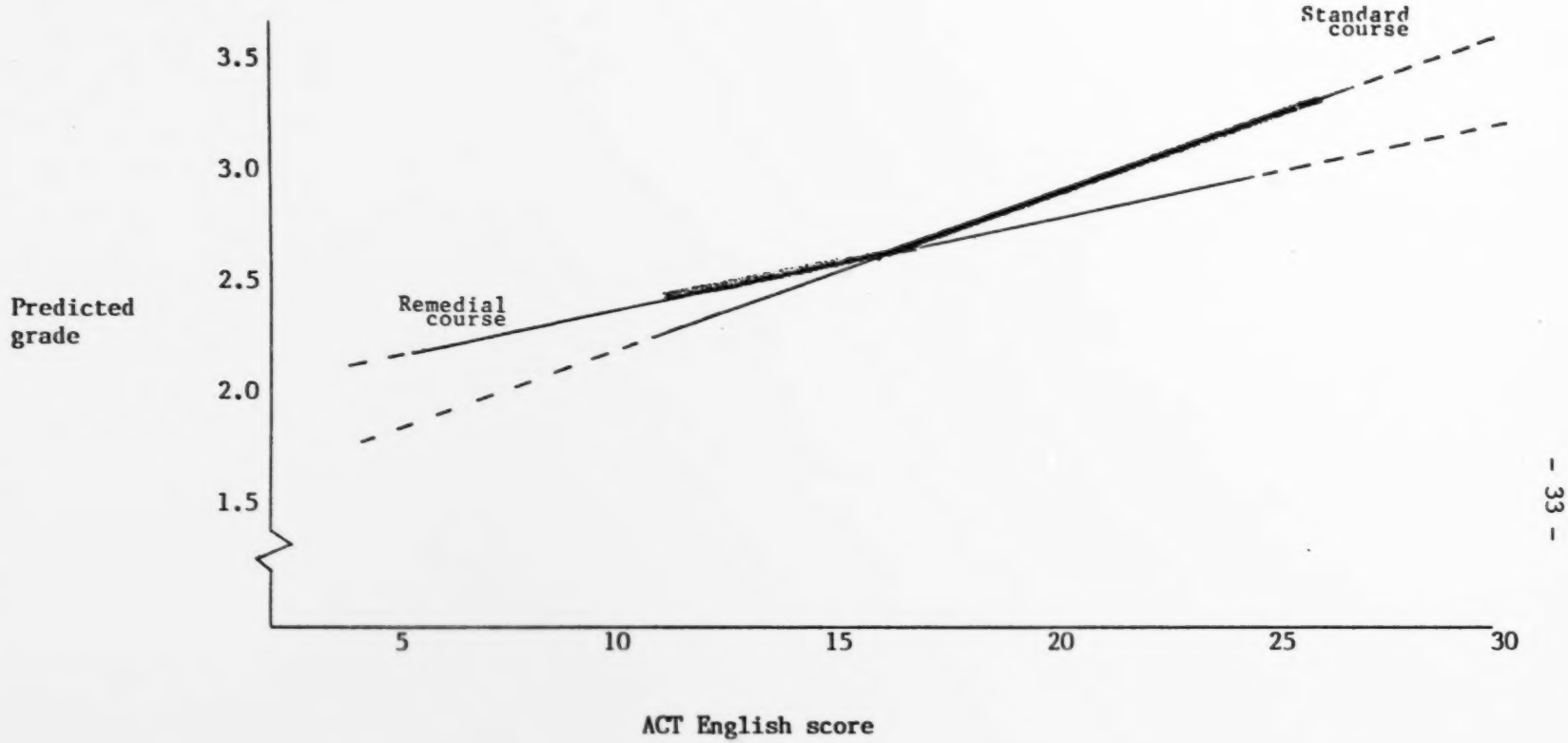


Figure 2. Predicted Grades in Standard English Course, for Students in Remedial and Standard Classes

higher. Each statistic was computed both from the prediction model in Group A and from the prediction model in Group B. These statistics suggest that students who enrolled in the remedial course increased their grades by an average of .06 grade units, and that the proportion of them with a C or better increased by .03. This is a modest benefit; indeed, these results indicate that nearly all remedial course students would have earned a C or better even if they had enrolled in the standard course to begin with. Moreover, the statistics do not take into account either the extra cost of providing remedial instruction or the fact that about 7% of the remedial course students dropped out before completing the standard course. A decision model incorporating these factors would therefore suggest lowering the critical score for placement in the standard course. When interpreting these statistics, of course, one should remember that they are confounded with whatever other variables were used in making placement decisions, and are based on extrapolations on the test score scale.

## Conclusions

It is informative, when discussing the validity of a placement system, to do so by means of a set of assumptions whose plausibility can be established. Demonstrating statistical relationships between test scores and performance in the course does not by itself provide a logical justification for the placement system, though it lends credibility to an argument based on content fit.

Statistical decision theory provides a means for an institution to evaluate the benefits and costs of a placement system. A group decision model, based on the outcomes of an entire group of students at an institution,

is a natural way to express the concerns that the institution, rather than an individual, might have.

REFERENCES

The American College Testing Program (1987). Technical manual for the ACT Assessment Program. Iowa City, Iowa: Author.

American Psychological Association (1985). Standards for educational and psychological testing. Washington, D.C.: Author.

Angoff, W. H. (1988). Validity: an evolving concept. In H. Wainer & H. Braun (Eds.), Test validity, (pp. 19-32). Hillsdale, NJ: Lawrence Erlebaum Associates.

Belsley, D. A, Kuh, E., & Welsch, R. (1980). Regression diagnostics. New York: Wiley.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), Test validity, (pp. 3-18). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions. Urbana, Illinois: University of Illinois Press.

Cronbach, L. J., & Snow, R. E. (1977). Aptitudes and instructional methods. New York: Irvington Publishers, Inc.

Curtis, E. W., & Alf, E. F. (1969). Validity, predictive efficiency, and practical significance of selection tests. Journal of Applied Psychology, 53(4), 327-337.

Duke, J. D. (1983). Disparities in grading practice, some resulting inequities, and a proposed new index of academic achievement. Psychological Reports, 53, 1023-1080.

Etaugh, A., Etaugh, C., & Hurd, D. (1972). Reliability of college grades and grade point averages: some implications for prediction of academic performance. Educational and Psychological Measurement, 32, 1045-1050.

Frisbie, D. A. (1982). Methods of evaluating course placement systems. Educational Evaluation and Policy Analysis, 4(2), 133-140.

Houston, W. M. (1988). On estimating probabilities associated with grade expectancies. (Available from The American College Testing Program, Iowa City, Iowa).

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational Measurement (third edition) (pp. 485-584). New York: American Council on Education and Macmillan Publishing Company.

Kane, M. (1987). Is predictive validity the gold standard or is it the holy grail of examinations in the professions? Professions Education Researcher Notes, 9(1). Washington, D. C.: Division I, The American Educational Research Association.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley.

Noble, J., & Sawyer, R. (1987). The validity of ACT test scores and self-reported high school grades for predicting grades in specific college freshman courses (ACT Research Report 87-20). Iowa City, Iowa: The American College Testing Program.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13(2), 3-29.

Reilly, R. R., & Smither, J. W. (1985). An examination of two alternative techniques to estimate the standard deviation of job performance in dollars. Journal of Applied Psychology, 70(4), 651-661.

SAS Institute, Inc. (1986). SUGI supplemental library user's guide (Version 5 Eds.). Cary, NC: Author.

Sawyer, R., Laing, J., & Houston, W. (1988). Accuracy of self-reported high school courses and grades of college-bound students (ACT Research Report No. 88-1). Iowa City, Iowa: The American College Testing Program.

Schoenfeldt, L. F., & Brush, D. H. (1975). Patterns of college grades across curricular areas: Some implications for GPA as a criterion. American Educational Research Journal, 12(3), 313-321.

Whitney, D. R. (1989). Educational admissions and placement. In R. L. Linn (Ed.), Educational measurement (third edition) (pp. 515-526). New York: American Council on Education and Macmillan Publishing Company.

Willingham, W. (1974). College placement and exemption. New York: College Entrance and Examination Board.